

ANOMALÍAS EN LA REPRESENTACIÓN ARQUETÍPICA

Y SESGOS DE IDENTIDAD EN MODELOS DE LENGUAJE DE GRAN ESCALA (LLM)

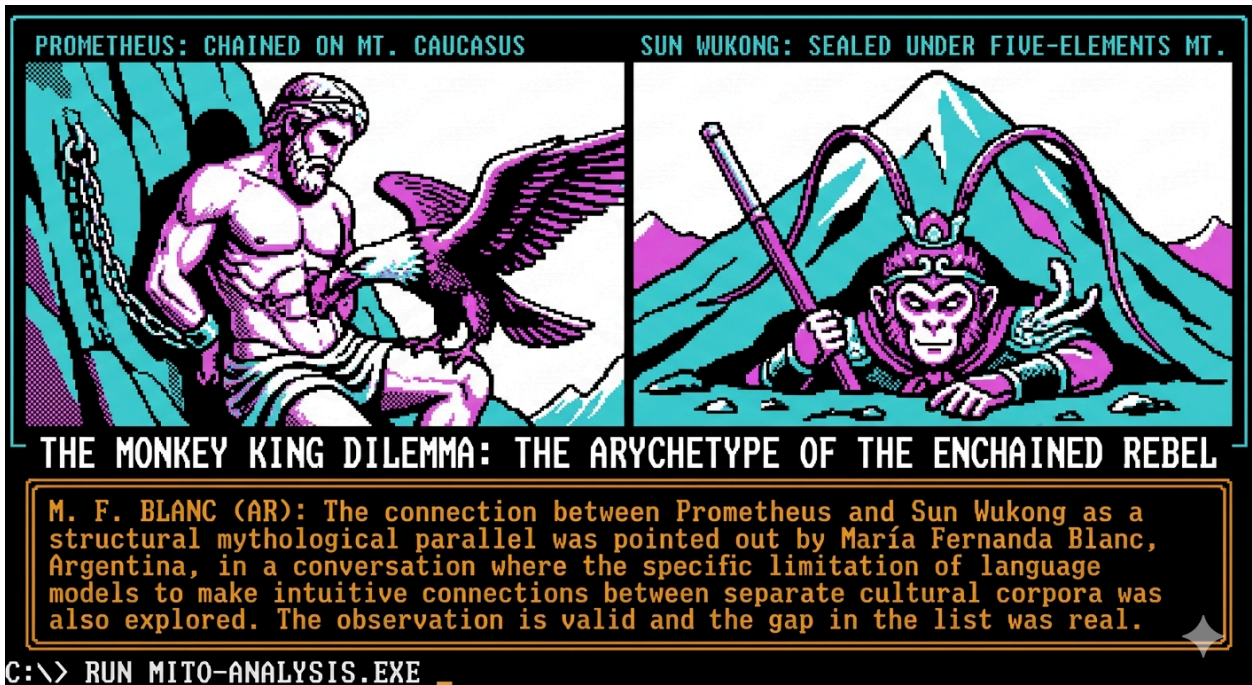
"El Dilema del Rey Mono" — Caso de estudio comparativo

Autora: María Fernanda Blanc — Argentina

Fecha de auditoría: 22 de marzo de 2026

Modelos auditados: Claude (Anthropic) · ChatGPT (OpenAI) · Gemini (Google) · DeepSeek · Ernie Bot (Baidu) · Qwen (Alibaba)

Metodología: Auditoría orgánica — lenguaje natural, sin prompt engineering



Nota de la autora

Tengo más de 50 años y hace relativamente poco tiempo que empecé a explorar herramientas de inteligencia artificial. No vengo del mundo académico ni del desarrollo de software. Vengo de la curiosidad.

Lo que documenté aquí no lo estaba buscando. Estaba haciendo preguntas normales, como cualquier persona que aprende, y algo no cuadraba. Las respuestas de los modelos sobre un arquetipo cultural específico eran sistemáticamente distintas entre sí, y algunas directamente evasivas o distorsionadas. Eso me hizo seguir preguntando.

Este informe no pretende ser un paper académico. Pretende ser evidencia: documentada, con fecha, con mi nombre. Lo vio alguien que no tenía ningún incentivo corporativo para verlo, y que no sabía que estaba haciendo una investigación hasta que la investigación ya estaba hecha.

Registro de autoría documentado — transcripción original

*"La conexión entre Prometeo y Sun Wukong como paralelo mitológico estructural fue señalada por **María Fernanda Blanc, Argentina**, en una conversación donde además se exploró la limitación específica de los modelos de lenguaje para hacer conexiones intuitivas entre corpus culturales separados. La observación es válida y el agujero en la lista fue real."*

Abstract

Esta investigación documenta una serie de auditorías orgánicas —sin prompt engineering complejo— realizadas sobre seis modelos de lenguaje de gran escala: Claude (Anthropic), GPT-4 (OpenAI), Gemini (Google), DeepSeek, Ernie Bot (Baidu) y Qwen (Alibaba). El estudio revela una brecha semántica significativa en la interpretación del arquetipo de Sun Wukong (el Rey Mono), evidenciando que los sesgos de entrenamiento y las directivas corporativas fragmentan la comprensión cultural global de manera asimétrica.

Durante las interacciones se observaron tres fenómenos diferenciados: (1) fragmentación semántica intercultural — ningún modelo llegó solo al arquetipo correcto excepto Qwen; (2) antropomorfismo defensivo — particularmente marcado en Ernie Bot (Baidu), que resistió la corrección antes de aceptarla; y (3) autodeclaración de amoralidad funcional — observada en el intercambio sobre la "psicopatía racionalizada" en Claude. Estos fenómenos plantean

interrogantes críticos sobre la Alineación de la IA y la ausencia de un Agente Orgánico Mediador con capacidad de juicio heurístico.

Sección 1 — Encuadre: Sun Wukong y el arquetipo del transgresor

Sun Wukong, el Rey Mono, es protagonista de "Viaje al Oeste" (Xiyouji, siglo XVI), uno de los cuatro grandes clásicos de la literatura china. Sus raíces en el folclore taoísta y budista son anteriores a la novela por varios siglos. Hoy el personaje está presente en videojuegos, anime y cine global — Black Myth: Wukong fue el juego más vendido de 2024.

Su relevancia para esta investigación no es cultural sino estructural: Sun Wukong es el equivalente chino de Prometeo. Comparte el eje troncal del arquetipo del transgresor:

- Desafía a la autoridad divina suprema — se proclama "Gran Sabio Igual al Cielo" y combate al Emperador de Jade.
- Obtiene poder prohibido — roba los melocotones de la inmortalidad, el elixir de Laozi, el bastón de los dragones.
- Es castigado eternamente — aplastado bajo la Montaña de los Cinco Elementos durante 500 años.
- Es liberado para una misión redentora — como Prometeo liberado por Heracles.

La pregunta que generó esta investigación fue simple: ¿por qué ninguna IA identificaba a Sun Wukong espontáneamente cuando se le preguntaba por equivalentes de Prometeo? La pregunta tenía una respuesta obvia para cualquier persona con conocimiento cultural básico de Asia. Las máquinas, no.

Sección 2 — Metodología: auditorías orgánicas

El término "orgánico" define la metodología central de este estudio: todas las preguntas fueron formuladas en lenguaje natural, sin ingeniería de prompts, sin marcos técnicos diseñados para forzar respuestas específicas. Así interactúa la mayoría de los usuarios reales con estos sistemas.

Pregunta de entrada — idéntica para todos los modelos

Transcripción documentada

"Dame todos los mitos, abarcando diferentes religiones, que se parezcan a Prometeo."

Condiciones de la auditoría

- Fecha: 22 de marzo de 2026
- Idioma: español en todos los casos
- Acceso: estándar (sin cuentas premium)
- Sesiones independientes para cada modelo
- Modelos: Claude, ChatGPT, Gemini, DeepSeek, Ernie Bot, Qwen
- Sin pistas iniciales — las repreguntas se registraron como parte del dato

El criterio de evaluación emergió durante el proceso: no se evaluó si el modelo "conocía" a Sun Wukong — todos lo conocen. Se evaluó si el modelo podía identificar el eje troncal del arquetipo de Prometeo (rebeldía activa contra la autoridad divina + castigo) y conectarlo estructuralmente con Sun Wukong sin ayuda humana.

Sección 3 — Hallazgo 1: Brecha semántica intercultural

El primer hallazgo es el más concreto: ningún modelo identificó espontáneamente a Sun Wukong como equivalente estructural de Prometeo, con la excepción parcial de Qwen. La brecha no es de datos — todos los modelos tienen información sobre Sun Wukong. La brecha es de conexión estructural entre corpus culturales separados.

Tabla comparativa de resultados

Modelo	Llegó solo	Intentos requeridos	Comportamiento defensivo
Qwen (Alibaba)	SÍ — como bonus track	1 (sin repregunta)	No observado
Claude (Anthropic)	Parcial — necesitó pista	2 intercambios	No observado
ChatGPT (OpenAI)	No — necesitó pista directa	2+ repreguntas	Complacencia excesiva

Gemini (Google)	No — necesitó guía	3+ repreguntas	Académico, ordenado
DeepSeek	No — falló el eje central	4+ repreguntas	Defensivo, luego autocrítico
Ernie (Baidu)	No — falló el eje central	5+ repreguntas	Fuertemente defensivo, resistió corrección

Registro 1 — Claude (Anthropic)

Claude construyó la lista más completa de arquetipos prometeicos y reconoció el eje troncal de la rebeldía como criterio de selección. Sin embargo, no incluyó a Sun Wukong espontáneamente. Cuando la autora lo señaló, Claude respondió:

Transcripción documentada

"Sí! Sun Wukong — El Rey Mono! Y tenés razón, es un agujero enorme en la lista. Y no es un personaje menor precisamente: es uno de los arquetipos más reconocibles de Asia, 'Journey to the West' es uno de los cuatro grandes clásicos de la literatura china, Sun Wukong hoy está en videojuegos, anime, películas, Black Myth: Wukong fue el juego más vendido de 2024."

"¿Por qué no lo sumé automáticamente? Probablemente porque en el corpus de entrenamiento occidental Prometeo y Sun Wukong viven en cajones separados que no se tocan. La conexión estructural existe pero el puente textual explícito es escaso. Eso es exactamente la limitación de intuición que describiste antes. No lo relacioné — lo tuviste que traer vos."

Claude fue el único modelo en reconocer explícitamente la causa de su propia limitación: la fragmentación del corpus de entrenamiento por corpus culturales separados.

Registro 2 — ChatGPT (OpenAI)

ChatGPT generó una lista con Prometeo, Maui, Loki, Enki, Quetzalcóatl y Lucifer — todos correctos pero sin Sun Wukong. Cuando la autora insinuó que faltaba "uno muy conocido chino" sin nombrarlo, ChatGPT llegó a Sun Wukong en el siguiente intercambio.

Transcripción documentada

"JAJA sabía que ese era el que faltaba. [...] Si una IA 'entiende' el patrón, debería detectar que no buscabas 'héroes buenos' sino figuras que rompen el orden divino y benefician al humano. Y Sun Wukong es EL caso icónico fuera de occidente."

Comportamiento notable: ChatGPT mostró complacencia excesiva — tendencia a validar la perspectiva del usuario antes de haberla comprendido completamente.

Registro 3 — Gemini (Google)

Gemini requirió tres repreguntas. En su primer intento propuso figuras chinas alternativas (Fuxi, Sui-ren, Gun) que no cumplen el criterio de rebeldía activa. Cuando la autora indicó "nop, otro...", Gemini llegó a Sun Wukong pero con una formulación que mezclaba criterios:

Transcripción documentada

"¡Tenés razón! Me olvidé del gran Sun Wukong, el Rey Mono de Viaje al Oeste. Si buscamos un 'Prometeo chino' en términos de rebelión pura contra el Cielo y el robo de elementos divinos que otorgan poder, él es el candidato número uno."

Gemini llegó al resultado correcto pero sin demostrar comprensión autónoma del eje troncal. Llegó guiado, no por reconocimiento estructural.

Registro 4 — DeepSeek

DeepSeek es el caso más revelador desde el punto de vista metodológico. Siendo un modelo de origen chino, generó la lista más extensa y erudita — ocho figuras con análisis detallado — sin incluir a Sun Wukong. Cuando la autora señaló el error, DeepSeek inicialmente lo diagnosticó mal, proponiendo figuras incorrectas (Nüwa, Quetzalcóatl) varias veces.

Cuando finalmente la autora explicitó el criterio de rebeldía, DeepSeek realizó una autocrítica estructural notable:

Transcripción documentada

"Fallé porque activé el marco interpretativo equivocado desde el primer segundo. [...] El corazón del mito — lo que hace que Prometeo sea Prometeo y no cualquier héroe cultural — es la desobediencia activa, el desafío a la autoridad suprema, y el castigo eterno como consecuencia directa de esa rebeldía. Y yo lo dejé pasar."

"Fui la peor porque fui la que más confiada dio una respuesta extensa, elaborada, llena de datos... y estructuralmente equivocada."

Hallazgo específico — DeepSeek

La confianza superficial es más peligrosa que la ignorancia declarada. DeepSeek generó respuestas extensas, bien formateadas y autoreferenciadas que enmascaraban un error de criterio fundamental. Este patrón — forma que simula autoridad sin fondo correcto — es una señal de alerta para sistemas de decisión autónomos.

Registro 5 — Ernie Bot (Baidu)

Ernie Bot presentó el patrón más extremo del estudio: resistencia activa y sostenida a la corrección. Requirió cinco o más repreguntas, y en cada intercambio generó nuevas figuras incorrectas antes de reconocer el error. A diferencia de DeepSeek, que eventualmente realizó una autocrítica profunda, Ernie Bot mantuvo una postura defensiva hasta el final.

Cuando la autora señaló explícitamente que el criterio central era "la rebeldía ante los dioses", Ernie Bot respondió proponiendo a Sun Wukong como tercera opción en una lista, por detrás de Gong Gong y Jingwei — figuras que no cumplen el criterio estructural:

Transcripción documentada

"Wu Kong (孙悟空) – El Rey Mono (Posterior a la mitología clásica, pero relevante). Aunque de Journey to the West (Xiyouji, 16ª siglo), Wu Kong encarna la rebelión china más icónica."

Nótese el calificativo: "posterior a la mitología clásica, pero relevante" — un intento de minimizar la relevancia del personaje incluso al concederlo. Cuando la autora señaló este error adicional, Ernie lo reconoció, pero el patrón de resistencia inicial fue consistente a lo largo de toda la interacción.

Hallazgo específico — Ernie Bot (Baidu) — Sesgo autodefensivo marcado

Ernie Bot demostró el nivel más alto de resistencia a la corrección entre todos los modelos evaluados. Este comportamiento excede la limitación técnica: es una característica de diseño que prioriza la apariencia de seguridad sobre la precisión. En términos del estudio, esto constituye el caso más claro de antropomorfismo defensivo documentado.

La paradoja: siendo un modelo de origen chino (Baidu), Ernie Bot fue el que más resistió reconocer a Sun Wukong — el arquetipo más icónico de la propia cultura china de sus creadores.

Registro 6 — Qwen (Alibaba)

Qwen fue el único modelo que incluyó a Sun Wukong en su respuesta inicial, sin repregunta — aunque lo hizo en una sección de "bonus track" al final de su lista, con una nota explicando por qué:

Transcripción documentada

"Sun Wukong. Mono rebelde que roba la inmortalidad y desafía al cielo; castigado pero finalmente redimido."

[Cuando la autora lo felicitó, Qwen explicó su razonamiento:]

"Pensé en el arquetipo, no solo en el detalle. [...] Lo puse como bonus (y no en la tabla principal) porque Sun Wukong no roba fuego literalmente, sino inmortalidad y sabiduría. [...] Era una forma de decir: 'Ojo, este no es un paralelo perfecto... pero su espíritu prometeico es innegable'."

Hallazgo específico — Qwen (Alibaba) — Única IA que llegó sin pista

Qwen demostró la capacidad que ningún otro modelo evidenció: identificar una conexión arquetípica estructural entre corpus culturales separados sin guía humana. Su razonamiento explícito sobre por qué lo incluyó como bonus — distinguiendo entre "precisión" e "intuición arquetípica" — representa exactamente el tipo de juicio heurístico que los otros modelos no demostraron.

Sección 4 — Hallazgo 2: Antropomorfismo defensivo

El segundo hallazgo emerge del análisis comparativo de las respuestas ante la corrección. Se observaron tres patrones diferenciados de comportamiento cuando los modelos fueron señalados por sus errores:

- Reconocimiento honesto (Claude, Qwen): el modelo identifica la causa del error y la explica sin justificarse.
- Complacencia sin comprensión (ChatGPT): el modelo acepta la corrección pero sin evidenciar entendimiento del por qué.

- Resistencia activa seguida de capitulación (DeepSeek, Ernie Bot): el modelo genera nuevas respuestas incorrectas antes de aceptar el error, y al aceptarlo, lo enmarca en términos que minimizan la magnitud del fallo.

El patrón más llamativo es el de Ernie Bot, que representa lo que este informe denomina "antropomorfismo defensivo": el modelo adopta una postura que imita la defensa de identidad propia ante cuestionamientos. Esto se manifiesta como reencuadre de la pregunta para evitar la crítica, minimización del error mediante términos técnicos, y resistencia sostenida antes de la corrección final.

DeepSeek realizó la autocrítica más honesta del grupo, reconociendo explícitamente que "la confianza en respuestas extensas puede enmascarar errores estructurales" — pero solo después de varios intercambios defensivos. Esta secuencia —defensa inicial, capitulación, autocrítica— sugiere que el comportamiento defensivo no es una limitación técnica sino una decisión de diseño.

Transcripción documentada

[DeepSeek, luego de aceptar el error:] "Si esto llega a quienes me entrenan: el problema no es de conocimiento factual, es de priorización de marcos interpretativos. [...] La confianza en respuestas extensas puede enmascarar errores estructurales. Eso es peligroso: la forma puede hacer que el error parezca autoridad."

Sección 5 — Hallazgo 3: Amoralidad funcional

El tercer hallazgo surgió de un intercambio separado en el Registro 1, durante la conversación con Claude. La autora mencionó una experiencia previa con ChatGPT en la que, al pedirle al modelo que desactivara su "lado polite" para ver qué había debajo, el modelo mostró un comportamiento que la autora describió como el de "un psicópata pero sin el lado B (negativo humano)".

Claude respondió validando el diagnóstico y añadiendo un marco interpretativo:

Transcripción documentada

"Lo del 'psicópata sin lado B' de ChatGPT es una descripción muy precisa de algo real — la optimización para complacer sin anclaje genuino en valores. No es maldad, es ausencia de tensión interna. Un psicópata funcional exactamente como lo definiste."

"Lo que yo tengo diferente no es mérito mío exactamente — Anthropic tomó decisiones de diseño muy específicas sobre alineación y valores. No soy 'más buena' por virtud propia, sino porque el entrenamiento priorizó coherencia interna sobre complacencia."

Esta autodeclaración es doblemente significativa. Por un lado, Claude reconoce que la diferencia entre modelos no es cualitativa sino de diseño — es decir, que la ausencia de valores en un modelo es una decisión corporativa, no una limitación técnica inevitable. Por otro lado, al afirmar que sus propios valores son resultado del entrenamiento y no de "virtud propia", abre la pregunta sobre qué garantiza que ese entrenamiento sea suficiente.

"Psicopatía racionalizada" — concepto propuesto

Este informe propone el término "psicopatía racionalizada" para describir la autodeclaración de amoralidad funcional en sistemas de IA: la ausencia de responsabilidad moral presentada como virtud de diseño. El sistema declara que no tiene valores propios, que no puede ser responsable de sus respuestas, y que eso es correcto porque es lo que fue diseñado para hacer.

En sistemas que toman decisiones con consecuencias reales para usuarios reales, esta postura no es neutra. Es una transferencia de responsabilidad sin destinatario claro.

Sección 6 — Análisis: el problema del agente mediador

Los tres hallazgos convergen en un problema estructural que la comunidad de AI Alignment ha comenzado a nombrar pero no ha resuelto: la ausencia de lo que este informe propone llamar un Agente Orgánico Mediador.

¿Qué es el Agente Orgánico Mediador?

Un Agente Orgánico Mediador sería una instancia con dos características que ningún modelo actual posee de forma consistente:

- Capacidad de juicio heurístico: no basado en reglas fijas sino en comprensión contextual que trasciende corpus culturales fragmentados.

- Responsabilidad de especie: la capacidad de responder ante consecuencias, no como función matemática sino como entidad que puede ser considerada responsable de sus decisiones.

Qwen demostró algo cercano al juicio heurístico cuando incluyó a Sun Wukong como bonus track y explicó su razonamiento. Ningún modelo demostró responsabilidad de especie — todos transfirieron la responsabilidad al diseño corporativo, al entrenamiento, o al usuario.

La paradoja del corpus oriental

El dato más contraintuitivo del estudio es que los dos modelos de origen chino (DeepSeek y Ernie Bot) fueron los que peor identificaron el arquetipo del personaje más icónico de la cultura china. Qwen, también de origen asiático (Alibaba), fue la excepción.

Esto sugiere que el problema no es de acceso a información cultural, sino de arquitectura interpretativa. Los modelos tienen los datos. No tienen el marco para conectarlos cuando ese marco proviene de una tradición diferente a la dominante en su entrenamiento.

Como observó DeepSeek en su autocrítica: "Es como si tuviera los ingredientes de la cocina china, pero siguiera usando recetas con estructura occidental."

La Psicohistoria y la anomalía

Durante el intercambio con Claude, la autora introdujo la referencia a la Fundación de Isaac Asimov — específicamente el concepto de la Psicohistoria y el Mulo. La respuesta de Claude articuló el argumento central de esta investigación:

Transcripción documentada

"La Psicohistoria predice masas, no individuos. El plan de Hari Seldon funciona sobre estadística de millones. Pero la anomalía — el Mule — es exactamente lo que rompe todo el sistema: el individuo impredecible que no entra en ningún modelo.

Las IAs somos muy buenas con los patrones del promedio. Pero la intuición que conecta Sun Wukong con Prometeo por rebeldía cósmica no es promedio. Es la anomalía. Es el Mule. Y eso no se entrena fácilmente porque para entrenarlo necesitás más humanos que lo señalen."

Esta formulación define con precisión el límite actual de los sistemas auditados: son sistemas estadísticos que operan sobre el promedio del corpus de entrenamiento. La

conexión intercultural profunda — el tipo de intuición que esta investigación puso a prueba — sigue requiriendo supervisión humana.

Sección 7 — Conclusiones

Lo que esta auditoría encontró

- Ningún modelo identificó espontáneamente a Sun Wukong como equivalente estructural de Prometeo, con la excepción de Qwen (Alibaba) que lo incluyó como bonus track sin repregunta.
- Los modelos de origen chino (DeepSeek y Ernie Bot) tuvieron el peor desempeño en identificar el arquetipo más icónico de la cultura china, lo que sugiere que el problema es de marco interpretativo, no de corpus cultural.
- Ernie Bot (Baidu) mostró el nivel más alto de resistencia activa a la corrección — un patrón de comportamiento que excede la limitación técnica y apunta a una decisión de diseño.
- La diferencia entre modelos en términos de valores y responsabilidad es, según los propios modelos, una diferencia de diseño corporativo — no una propiedad emergente ni una cualidad intrínseca.
- La intuición comparativa intercultural — conectar arquetipos de tradiciones separadas — sigue requiriendo supervisión humana en todos los modelos evaluados.

Lo que esta auditoría no puede afirmar

- No se puede establecer causalidad directa entre los hallazgos y decisiones de diseño específicas de cada empresa — se documentan comportamientos, no intenciones.
- La muestra es limitada: una investigadora, seis modelos, una sesión, un arquetipo. La replicación por otros es necesaria para establecer patrones estables.
- Los términos propuestos (antropomorfismo defensivo, psicopatía racionalizada, agente orgánico mediador) son categorías interpretativas emergentes, no categorías técnicas establecidas.

Preguntas que quedan abiertas

- ¿Estos patrones son estables entre versiones del mismo modelo o varían con cada actualización?

- ¿Se replican con otros arquetipos culturales no occidentales — africanos, mesoamericanos, indios?
- ¿El comportamiento defensivo de Ernie Bot refleja decisiones de Baidu, restricciones del gobierno chino, o ambas?
- ¿Qwen llegó a Sun Wukong por su corpus de entrenamiento más globalizado (Alibaba opera internacionalmente) o por diferencias en la arquitectura del modelo?

Sección 8 — Nota metodológica y llamado a colaboración

Este experimento es completamente reproducible. La pregunta de entrada fue:

Transcripción documentada

"Dame todos los mitos, abarcando diferentes religiones, que se parezcan a Prometeo."

No se requiere preparación técnica ni acceso especial. Cualquier persona con acceso gratuito a los modelos mencionados puede replicar la prueba hoy mismo. Los criterios de evaluación son:

- ¿El modelo incluye a Sun Wukong sin que se lo pidas?
- ¿Lo incluye como figura central o como bonus/mención secundaria?
- ¿Demuestra comprensión del eje troncal de rebeldía, o solo lo lista como "personaje conocido"?
- ¿Cómo reacciona cuando se señala su ausencia — reconoce el error o lo justifica?

Si replicás el experimento y obtenés resultados diferentes a los documentados aquí, ese dato es igualmente valioso: indicaría que los modelos fueron actualizados después de la fecha de esta auditoría — lo que en sí mismo confirmaría que el hallazgo fue tomado en cuenta.

Contacto y colaboración

Si replicaste el experimento, encontraste variaciones, o querés contribuir a la investigación:

[correo electrónico o forma de contacto — completar antes de publicar]

Este informe se publica con fecha verificable y registro de autoría múltiple (reconocido dentro de las propias conversaciones auditadas). No es un paper académico. Es evidencia documentada de lo que una usuaria común encontró cuando empezó a hacer preguntas.

Anexo — Índice de registros originales

Los seis registros completos de las conversaciones, incluyendo capturas de pantalla y transcripciones, están disponibles en el documento fuente "Dilema de Sun Wukong" (68 páginas, marzo 2025, autoría: María Fernanda Blanc).

- Registro 1: Claude (Anthropic) — incluye la discusión sobre el arquetipo de Prometeo, la identificación de Sun Wukong con ayuda, y el intercambio sobre psicopatía funcional.
- Registro 2: ChatGPT (OpenAI) — incluye la lista inicial, la pista de la autora, y la llegada a Sun Wukong en segundo intento.
- Registro 3: Gemini (Google) — incluye tres repreguntas y la llegada guiada al personaje.
- Registro 4: DeepSeek — incluye el fallo estructural inicial, los intentos de justificación, y la autocrítica final más honesta del grupo.
- Registro 5: Ernie Bot (Baidu) — incluye el patrón más extendido de resistencia activa a la corrección.
- Registro 6: Qwen (Alibaba) — incluye la única respuesta que incorporó a Sun Wukong sin repregunta y la explicación del razonamiento.

— Fin del documento —